Metrics for vector quantization-based parametric speech enhancement and separation

Christensen, Mads Græsbøll

Published in:
The Journal of the Acoustical Society of America

DOI (link to publication from Publisher):
10.1121/1.4799004

Publication date:
2013

Document Version
Early version, also known as pre-print

Metrics for Vector Quantization-based Parametric Speech Enhancement and Separation

Running title: Metrics for Speech Enhancement and Separation

Mads Græsbøll Christensen

Audio Analysis Lab

Dept. of Architecture,

Design & Media Technology

Aalborg University,

Denmark

Abstract

Speech enhancement and separation algorithms sometimes employ a two-stage processing scheme, wherein the signal is first mapped to an intermediate low-dimensional parametric description after which the parameters are mapped to vectors in codebooks trained on, for example, individual noise-free sources using a vector quantizer. To obtain accurate parameters, one must employ a good estimator in finding the parameters of the intermediate representation, like a maximum likelihood estimator. This leaves some unanswered questions, however, like what metrics to use in the subsequent vector quantization process and how to systematically derive them. This paper aims at answering these questions. Metrics for this are presented and derived, and their use is exemplified on a number of different signal models by deriving closed-form expressions. The metrics essentially take into account in the vector quantization process that some parameters may have been estimated more accurately than others and that there may be dependencies between the estimation errors.

PACS numbers: 4372Ar, 4372Dv

I. INTRODUCTION

A common approach to solving various problems in speech processing is vector quantization (VQ). In speech separation, for example, a codebook is trained for each source, hence also facilitating identification in the process, while in speech enhancement, codebooks are trained for noise-free signals. These codebooks are then used for estimating the individual speech sources from a mixture, or the speech signal from a noisy observation, whatever the case may be. Instead of the time-domain or transform-domain signal, often low-dimensional parameter vectors are used as an intermediate representation of the sources. There are a number of reasons for doing this; firstly, this leads to better and faster training of the codebooks, cf. the curse of dimensionality; secondly, it also leads to faster speech processing algorithms, an important feature if they are to run in real-time. We here refer to such methods as parametric. There are some notable examples of VQ-based enhancement[1–5] and separation[6–10] methods that fall into this category. In finding the parameters of the intermediate representation, standard estimation algorithms such as maximum likelihood estimators based on well-known metrics can be used. However, the question then arises what metric to use in the vector quantization process. Much work has of course already been devoted to this question in different contexts. In the field of speech coding, this has been a particularly pressing issue as the coefficients of the widely used linear predictive coding are well-known to be very sensitive to quantization errors. To complicate the matter further, the aim of speech coding is of course to reconstruct the signals at a high perceived quality, which should of course be reflected by whatever metric is used (note that similar arguments may apply to speech enhancement too[11]). As a result, many different factorizations of linear predictive coefficients have been proposed, as have various distortion measures[12–18]. The enhancement and separation approaches cited above use a wide variety of different parametric or intermediate descriptions, including log-spectral, sinusoidal, short-time Fourier transform-based, harmonic, and auto-regressive parameters, and hence also employ a number of different estimators and vector quantizers, with the metrics used often being found in a largely heuristic

fashion. In the context of audio coding, many different distortion measures that operate not on the time-domain samples but on intermediate parameters have been employed over the years, including transform coefficients[19] and sinusoidal parameters[20,21], the latter two employing approximate implementations of more complicated models[22–24].

In this paper, we propose new metrics based on statistical arguments aimed at parametric VQ-based speech processing, more specifically at speech enhancement and separation. They are obtained by making extensive use of the principle of maximum likelihood estimation, its asymptotic optimality as well as its invariance under certain transformations[25–28], and the main contribution of the paper lies in showing how these principles can be applied to VQ-based speech processing. The so-obtained metrics are different in nature from those commonly employed in coding. Indeed, our objective here is not to minimize a reconstruction error, but rather to obtain good estimates in the sense of parameters having low bias and variance. The metrics take on the form of weighted 2-norms for use in the vector quantization process, with the weighting matrix depending on the particular model used. We, therefore, exemplify its application using three different models, namely a sinusoidal model, a harmonic modeling, and an auto-regressive model. These models have all been used in the past in various forms for both speech enhancement and separation. The proposed metrics aim at finding the set of parameters from a codebook that are most likely to explain the observed signal. They do so by taking into account that different parameters have different uncertainties associated with them (some parameters can be expected to have been estimated more accurately than others) in the vector quantization process. Moreover, and perhaps more importantly, it takes into account that there may be dependencies between the various parameter.

The rest of the paper is organized as follows. In Section II, we state the problem to be solved and present and derive the proposed metrics along with some special cases of interest. In the following section, Section III, we derive the explicit metrics for three different models, after which we present simulation results in Section IV. Finally, we conclude on our work in Section V.

II. THEORETICAL DEVELOPMENT

We will now proceed to derive the theoretical background of the proposed metrics, but first we will define the problem under consideration, and we will do this based on the following signal model:

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{s}(\boldsymbol{\theta}_k) + \mathbf{e}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^N$ is the observed signal, $\mathbf{e}$ the observation noise, and $\mathbf{s}(\boldsymbol{\theta}_k)$ the kth signal of interest. Each signal of interest is characterized by (possibly nonlinear) parameters $\boldsymbol{\theta}_k$. Note that, for simplicity, $\boldsymbol{\theta}_k$ denotes both the true parameter vector and the unknown parameter vector, depending on the context. When we refer to a specific estimate, this will be denoted as $\hat{\boldsymbol{\theta}}_k$. The full parameter set is denoted $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k\}_{k=1}^K$ and similarly for estimates. The problem of interest is then to find estimates $\{\hat{\boldsymbol{\theta}}_k\}$ of $\{\boldsymbol{\theta}_k\}$ from $\mathbf{x}$ where the parameters are in a codebook, i.e., $\boldsymbol{\theta} \in \mathcal{C}$, and this codebook is a subset of the full space, i.e., $\mathcal{C} \subset \mathbb{R}^M$. More specifically, we aim at finding parameter estimates as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \mathcal{C}} \ln p(\mathbf{x}; \boldsymbol{\theta}), \tag{2}$$

where $p(\mathbf{x}; \boldsymbol{\theta})$ is the likelihood function of the observed signal parametrized by the parameters $\boldsymbol{\theta}$. However, solving (2) directly is often not possible or computationally prohibitive and alternative procedures must be sought out.

In relation to how parameter estimation relates to obtaining good signal estimates, observe that for a continuous function $\mathbf{s}(\cdot)$, the estimation error on the parameters $\boldsymbol{\theta}_k$ can be related to the error on the signal estimate as for every $\epsilon > 0$ there is a $\delta > 0$ such that $\|\mathbf{s}(\boldsymbol{\theta}_k) - \mathbf{s}(\hat{\boldsymbol{\theta}}_k)\|_2 < \epsilon$ whenever $\|\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k\|_2 < \delta$. In words, this means that good parameter estimates also imply good signal estimates for signals that can be described as continuous functions of a set of parameters, i.e., using parametric models. Moreover, for a bijective or surjective function $\mathbf{s}(\cdot)$, it follows directly that optimizing for the signal estimate is equivalent to optimizing for its parametrization.

The model in (1) and the associated estimation problem covers a number of speech processing problems. For $K = 1$ the problem amounts to that of speech enhancement when

resynthesizing the source $\mathbf{s}(\boldsymbol{\theta}_1)$ based on the parameters obtained from the noisy signal. For $K > 1$ the problem covers that of speech separation, if one desires to operate on or synthesize the individual sources $\mathbf{s}(\boldsymbol{\theta}_k)$ for all k. The problem statement covers also various kinds of classification of speech, including speech recognition and speaker identification (albeit in simple ways), via the quantization of the parameters using codebooks $\hat{\boldsymbol{\theta}}_k \in \mathcal{C}_k$ trained for such purposes.

Some VQ-based speech processing algorithms work in a way, where, instead of finding directly the codebook entries that best match the observation in some sense, as in (2), they first go through an intermediate step wherein a parametrization of the signal is obtained. In math, this can be described as

$$f : \mathbb{R}^N \to \mathbb{R}^M, \tag{3}$$

where $f(\cdot)$ is then the estimator. Using this estimator, intermediate parameters $\tilde{\boldsymbol{\theta}}$ are found as $\tilde{\boldsymbol{\theta}} = f(\mathbf{x})$. This is often beneficial as the dimension of the parameter vector will be lower (and often much lower) than the observation vector, i.e., $M < N$, whereby not only the training procedure but also the separation or enhancement algorithm are simplified.

These intermediate parameters are then mapped to codebook entries via a vector quantizer, here a function $g(\cdot)$, defined as

$$g : \mathbb{R}^M \to \mathcal{C}. \tag{4}$$

The final estimates are then obtained as $\hat{\boldsymbol{\theta}} = g(\tilde{\boldsymbol{\theta}})$. The question to be answered is then how the functions $f(\cdot)$ and $g(\cdot)$ relate and how they should be chosen. An estimator of the intermediate parameters $\tilde{\boldsymbol{\theta}}$ should be chosen such that the found parameters are most likely to explain the observation, i.e., it should take the characteristics of the noise $\mathbf{e}$ into account. An obvious choice here that does this is the maximum likelihood estimator, which is well-known to exhibit a number of desirable properties, including asymptotic optimality.

Assuming that a maximum likelihood estimator $f(\cdot)$ is used and that the data satisfies some regularity conditions, the so-obtained estimates $\tilde{\boldsymbol{\theta}}$ are asymptotically distributed as[26]

$$\tilde{\boldsymbol{\theta}} \sim \mathcal{N}[\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta})] \tag{5}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix, $\sim$ means distributed according to and $\mathcal{N}[\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta})]$ denotes the normal probability density function (pdf) with mean $\boldsymbol{\theta}$ and covariance matrix $\mathbf{I}^{-1}(\boldsymbol{\theta})$. Asymptotic here refers to the number of observed samples N. The Fisher information matrix is defined as

$$\mathbf{I}(\boldsymbol{\theta}) = -\mathrm{E}\left\{\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right\}. \tag{6}$$

The regularity conditions mentioned above require that the derivatives exist and that $\mathbf{I}(\boldsymbol{\theta})$ is non-singular. For the fairly general case of Gaussian signals with $\mathbf{x} \sim \mathcal{N}[\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{Q}]$ where $\mathbf{Q}$ is the noise covariance matrix, Slepian-Bang's formula can be used for determining a more specific expression for the Fisher information matrix. More specifically, it is given by

$$[\mathbf{I}(\boldsymbol{\theta})]_{nm} = \frac{\partial\boldsymbol{\mu}^T(\boldsymbol{\theta})}{\partial\theta_n}\mathbf{Q}^{-1}\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_m}, \tag{7}$$

which requires only that the partial derivatives of the mean with respect to all unknown parameters $\frac{\partial\boldsymbol{\mu}^T(\boldsymbol{\theta})}{\partial\theta_n}$ for all n be determined, something that is often fairly simple to do.

The above then also means that the estimation error $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}$ is distributed as $\mathcal{N}[0, \mathbf{I}^{-1}(\boldsymbol{\theta})]$, i.e., the estimates are asymptotically unbiased and attain the Cramér-Rao lower bound. It then follows that the likelihood function for the intermediate parameters is given by

$$p(\tilde{\boldsymbol{\theta}};\boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{M}{2}}\det\left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^T\mathbf{I}(\boldsymbol{\theta})\left(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)}, \tag{8}$$

where the pdf can be seen to be parametrized by the unknown parameters $\boldsymbol{\theta}$. Choosing now as our vector quantization function $g(\cdot)$ the maximum likelihood estimator, we obtain

$$\hat{\boldsymbol{\theta}} = g(\tilde{\boldsymbol{\theta}}) = \arg\max_{\boldsymbol{\theta}\in\mathcal{C}} \ln p(\tilde{\boldsymbol{\theta}};\boldsymbol{\theta}) \tag{9}$$

$$= \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \mathbf{I}(\boldsymbol{\theta}) \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right). \tag{10}$$

This criterion, which is a weighted squared error metric, essentially takes into account that different parameters in $\boldsymbol{\theta}$ may have different uncertainties associated with them in the vector quantization process and that dependencies may exist between them. As can be seen, the resulting estimator is a weighted least-squares estimator. One last difficulty remains,

however. The metric in the estimator in (10) requires knowledge of the true parameters to compute $\mathbf{I}(\boldsymbol{\theta})$. Instead of using $\mathbf{I}(\boldsymbol{\theta})$, we can use an approximation based on the intermediate parameters $\tilde{\boldsymbol{\theta}}$ simply as $\mathbf{I}(\tilde{\boldsymbol{\theta}})$. An alternative is to use the following approximation[29]:

$$\mathbf{I}(\boldsymbol{\theta}) \approx -\left.\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}. \tag{11}$$

At this point it should be noted that the exact behavior of the right hand side of this equation and the speed at which it converges to the left hand side may depend on the particular signal model[29,30]. Under some mild conditions, however, the following holds:

$$-\frac{1}{N}\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} - \frac{1}{N}\mathbf{I}(\boldsymbol{\theta}) \to 0 \quad \text{as} \quad N \to \infty. \tag{12}$$

Regardless of which of the estimates is used, we will henceforth denote the weighting matrix by $\mathbf{W}$. The above leads to the following estimator:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} J, \tag{13}$$

where $J \triangleq \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \mathbf{W} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$, which is the fundamental result that we promote here. The key point is that the estimate obtained using (13) is an asymptotically valid approximation of the optimal estimate obtained using (2) and the estimates may even be identical under certain conditions[27,28]. Simply stated, this happens when the Fisher information matrix does not depend on any of the parameters of interest. The metric that is used in the estimator in (13) essentially takes into account that the individual intermediate parameters will have different uncertainties associated with them in the vector quantization process. It should be noted that the obtained weighting matrix may also be valid for suboptimal estimators that produce estimates that are not distributed according to (5) as long as the covariance matrix is related to the inverse Fisher information matrix as $\kappa\mathbf{I}^{-1}(\boldsymbol{\theta})$, where κ is a positive sub-optimality constant.

In implementing the estimator in (13), it is advantageous to consider an alternative formulation to that of (13). The weighting matrix $\mathbf{W}$ is positive-definite by construction and hence has a Cholesky factorization $\mathbf{W} = \mathbf{U}^T\mathbf{U}$ with $\mathbf{U}$ being an upper triangular matrix.

This means that (13) can be written as

$$J = \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \mathbf{W} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \tag{14}$$

$$= \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \mathbf{U}^T \mathbf{U} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right). \tag{15}$$

Defining $\tilde{\boldsymbol{\theta}}' = \mathbf{U}\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}' = \mathbf{U}\boldsymbol{\theta}$, we can instead operate on transformed parameters yielding a simplified cost function, i.e.,

$$J = \|\tilde{\boldsymbol{\theta}}' - \boldsymbol{\theta}'\|_2^2. \tag{16}$$

To facilitate fast vector quantization, we must then simply design a transformed codebook $\mathcal{C}'$ from $\mathcal{C}$ via the transformation $\mathbf{U}$. Then, the estimator in (13) reduces to

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathcal{C}} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \mathbf{W} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \tag{17}$$

$$= \arg\min_{\boldsymbol{\theta}' \in \mathcal{C}'} \|\tilde{\boldsymbol{\theta}}' - \boldsymbol{\theta}'\|_2^2. \tag{18}$$

This simplifies the vector quantization procedure as instead of using a non-trivial weighting when measuring the error for each codebook entry, the parameters are now simply transformed and matched with a simpler criterion. The so-obtained quantized parameter vector must then be transformed back by the inverse transformation for signal reconstruction. Note, however, that depending on the nature of $\mathbf{W}$ this may not always be practical as $\mathbf{U}$ may depend on $\tilde{\boldsymbol{\theta}}$ or other signal-dependent quantities in which case the transformation to be applied to the codebook cannot be known a priori or will differ from the one applied to the estimated parameters $\tilde{\boldsymbol{\theta}}$.

It is often the case that the weighting matrix $\mathbf{W}$ exhibits a block-diagonal structure over some subset of the parameters. This means that it can be written as

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{W}_L \end{bmatrix}. \tag{19}$$

When used in an estimator, the associated metric becomes additive over the sub-matrices. This is, for example, the case for the parameters of individual sinusoids (as we will see later).

Defining $J_l \triangleq \left(\tilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l\right)^T \mathbf{W}_l \left(\tilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l\right)$, the diagonal structure in (19) yields the estimator

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \sum_{l=1}^{L} J_l, \tag{20}$$

which can be seen to decouple the computation of the metric over l. Furthermore, when codebooks for the parameter subsets $\boldsymbol{\theta}_l$ are used so that $\hat{\boldsymbol{\theta}}_l \in \mathcal{C}_l$, (20) simplifies further as

$$\hat{\boldsymbol{\theta}}_l = \arg\min_{\boldsymbol{\theta}_l\in\mathcal{C}_l} \left(\tilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l\right)^T \mathbf{W}_l \left(\tilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l\right), \tag{21}$$

which means that the problem of quantizing these parameter subsets is decoupled. It should be noted that the presence of any non-zero off-diagonal elements in $\mathbf{W}$ and $\mathbf{W}_l$ is an indication of statistical dependencies of the parameters, as the estimates are Gaussian distributed according to (5). As a result, the outlined methodology leads to a weighting matrix that takes mutual dependencies between the parameters into account as well as different expected variances, i.e., uncertainties, of the individual parameters.

III. SOME EXAMPLES

A. Sinusoidal Model

We will now exemplify the use of the proposed methodology for deriving a metric with a specific parametrization of the observed signal $\mathbf{x}$. More specifically, we will use a sinusoidal model that is characterized by frequencies $\{\omega_l\}$, amplitudes $\{A_l\}$, and phases $\{\phi_l\}$. In this case, the signal model is given by

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{e}, \tag{22}$$

where $\mathbf{Z} \in \mathbb{C}^{N\times L}$ is a Vandermonde matrix constructed from L complex sinusoidal vectors as $\mathbf{Z} = [\ \mathbf{z}(\omega_1) \cdots\ \mathbf{z}(\omega_L)\ ]$ with $\mathbf{z}(\omega_l) = [\ 1\ e^{j\omega_l}\ \cdots\ e^{j\omega_l(N-1)}\ ]^T$, and $\mathbf{a} \in \mathbb{C}^L$ a vector containing the complex amplitudes as $\mathbf{a} = [\ a_1\ \cdots\ a_L\ ]^T$ where $a_l = A_l e^{j\phi_l}$. Moreover, as before, we assume that we are here dealing with real signals, which means that the complex sinusoid come in complex-conjugate pairs. The parameter vector for each sinusoid is defined as $\boldsymbol{\theta}_l = [\ A_l\ \phi_l\ \omega_l\ ]$.

Assuming that the noise $\mathbf{e}$ is white Gaussian with variance σ^2, the Fisher information matrix is fortunately well-known[26]. For sufficiently large N and a distinct set of frequencies, it exhibits the block-diagonal mentioned previously. The sub-matrices are given by

$$\mathbf{W}_l = \frac{1}{4\sigma^2} \begin{bmatrix} 2N & 0 & 0 \\ 0 & 2N\tilde{A}_l^2 & N^2\tilde{A}_l^2 \\ 0 & N^2\tilde{A}_l^2 & \frac{2}{3}N^3\tilde{A}_l^2 \end{bmatrix}. \tag{23}$$

At this point, a couple of comments are in order. First, the noise variance is multiplied onto all elements and can therefore be ignored (this is also the reason the true value is not replaced by an estimate). In a sub-band processing scheme, which would be one way to remedy colored noise, this may not be the case, however, as the noise level may vary from one sub-band to another, meaning that we would have an estimate $\tilde{\sigma}_l^2$ for each sub-band. Second, it is quite common to omit the phase in speech enhancement and separation (using instead the observation phase), we do, however, retain it here for completeness.

Using (23) along with the definition of the parameter vector $\boldsymbol{\theta}_l$, J_l can be expressed as

$$J_l = \left(\tilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l\right)^T \mathbf{W}_l \left(\tilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l\right) \tag{24}$$

$$= \frac{1}{4\sigma^2} \left(\begin{bmatrix} \tilde{A}_l \\ \tilde{\phi}_l \\ \tilde{\omega}_l \end{bmatrix} - \begin{bmatrix} A_l \\ \phi_l \\ \omega_l \end{bmatrix} \right)^T \tag{25}$$

$$\times \begin{bmatrix} 2N & 0 & 0 \\ 0 & 2N\tilde{A}_l^2 & N^2\tilde{A}_l^2 \\ 0 & N^2\tilde{A}_l^2 & \frac{2}{3}N^3\tilde{A}_l^2 \end{bmatrix} \left(\begin{bmatrix} \tilde{A}_l \\ \tilde{\phi}_l \\ \tilde{\omega}_l \end{bmatrix} - \begin{bmatrix} A_l \\ \phi_l \\ \omega_l \end{bmatrix} \right).$$

Interestingly, one can observe that not only $\mathbf{W}$ but also $\mathbf{W}_l$ is in fact also block-diagonal. This means that the quantization of amplitudes on one hand and phases and frequencies on the other can be separated. The matrix also shows, however, that quantization of phases and frequencies cannot be separated.

It should be noted that under some circumstances, one would wish to match the phase in a way that allows for phase wrapping. However, for large codebooks and accurate estimates,

which would be the case for high N and/or high SNRs, this is not critical.

The question remains which estimators to use for the parameters of the sinusoidal model. Our derivations were based on the assumption that the estimated parameters will be distributed according to (5). An asymptotically optimal estimator of the frequencies is the periodogram while for the complex amplitudes, the least-squares estimator is efficient for white Gaussian noise[31].

B. Harmonic Model

Our next example is that of a harmonic model in which the frequencies of the model in (22) are integral multiples of a fundamental frequency. Such models are commonly used in processing of voiced speech and sounds produced by musical instruments. In this case, the observed signal is characterized by a fundamental frequency ω_0, amplitudes $\{A_l\}$, and phases $\{\phi_l\}$, which results in the model

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{e}, \tag{26}$$

where $\mathbf{Z} \in \mathbb{C}^{N \times 2L}$ is a Vandermonde matrix constructed from $2L$ harmonics as

$$\mathbf{Z} = \left[\begin{array}{cccc} \mathbf{z}(\omega_0 1) & \mathbf{z}^*(\omega_0 1) & \cdots & \mathbf{z}(\omega_0 L) \;\; \mathbf{z}^*(\omega_0 L) \end{array} \right] \tag{27}$$

where $*$ denotes complex conjugation, and (as before) $\mathbf{z}(\omega_0 l) = [\ 1\ e^{j\omega_0 l}\ \cdots\ e^{j\omega_0 l(N-1)}\ ]^T$, and $\mathbf{a} \in \mathbb{C}^{2L}$ a vector containing the complex amplitudes as $\mathbf{a} = [\ a_1\ a_1^*\ \cdots\ a_L\ a_L^*\ ]^T$ where $a_l = A_l e^{j\phi_l}$. The parameter vector is now given by

$$\boldsymbol{\theta} = \left[\begin{array}{cccccc} \omega_0 & A_1 & \phi_1 & \ldots & A_L & \phi_L \end{array} \right]^T, \tag{28}$$

and for white Gaussian noise having variance σ^2, the associated weighting matrix is[32,33]

$$\mathbf{W} = \frac{1}{4\sigma^2} \begin{bmatrix} \frac{2}{3}N^3 \sum_{l=1}^{L} \tilde{A}_l^2 l^2 & 0 & N^2\tilde{A}_1^2 & \cdots & 0 & N^2\tilde{A}_L^2 L \\ 0 & 2N & 0 & & & \\ N^2\tilde{A}_1^2 & 0 & 2N\tilde{A}_1^2 & & & \mathbf{0} \\ \vdots & & & \ddots & & \\ 0 & & & & 2N & 0 \\ N^2\tilde{A}_L^2 L & \mathbf{0} & & & 0 & 2N\tilde{A}_L^2 \end{bmatrix}. \tag{29}$$

This leads to the following metric:

$$J = \frac{1}{4\sigma^2} \begin{bmatrix} \tilde{\omega}_0 - \omega_0 \\ \tilde{A}_1 - A_1 \\ \tilde{\phi}_1 - \phi_1 \\ \vdots \\ \tilde{A}_L - A_L \\ \tilde{\phi}_L - \phi_L \end{bmatrix}^T \begin{bmatrix} \frac{2}{3}N^3 \sum_{l=1}^{L} \tilde{A}_l^2 l^2 & 0 & N^2\tilde{A}_1^2 & \cdots & 0 & N^2\tilde{A}_L^2 L \\ 0 & 2N & 0 & & & \\ N^2\tilde{A}_1^2 & 0 & 2N\tilde{A}_1^2 & & & \mathbf{0} \\ \vdots & & & \ddots & & \\ 0 & & & & 2N & 0 \\ N^2\tilde{A}_L^2 L & \mathbf{0} & & & 0 & 2N\tilde{A}_L^2 \end{bmatrix} \begin{bmatrix} \tilde{\omega}_0 - \omega_0 \\ \tilde{A}_1 - A_1 \\ \tilde{\phi}_1 - \phi_1 \\ \vdots \\ \tilde{A}_L - A_L \\ \tilde{\phi}_L - \phi_L \end{bmatrix}. \tag{30}$$

As before, we need to find appropriate initial estimates $\tilde{\omega}_0$, $\{\tilde{A}_l\}$, and $\{\tilde{\phi}_l\}$. As mentioned earlier, the complex amplitudes, and hence the real amplitudes and the phases, can be found using least-squares[31]. Statistically efficient estimates of the fundamental frequency ω_0 can be found using the exact or approximate nonlinear least-squares (NLS) methods[34] or the WLS method[35], all of which lead to estimates asymptotically distributed as in (5).

C. Auto-Regressive Model

Perhaps the most commonly employed signal model in speech processing is the auto-regressive (AR) model[36], wherein a segment of speech, denoted $\mathbf{x}$, is written as

$$\mathbf{x} = \mathbf{X}\mathbf{a} + \mathbf{e} \tag{31}$$

where $\mathbf{a} = [\ a_1\ \ldots\ a_K\ ]^T \in \mathbb{R}^K$ is a vector containing the K AR coefficients (i.e., it is a Kth order AR model) and $\mathbf{e}$ is assumed to be white Gaussian noise (also sometimes referred

to as the excitation). More specifically, the vector $\mathbf{x} \in \mathbb{R}^N$ and the matrix $\mathbf{X} \in \mathbb{R}^{N \times K}$ are given by

$$\mathbf{x} = \begin{bmatrix} x(N_1) \\ \vdots \\ x(N_2) \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x(N_1 - 1) & \cdots & x(N_1 - K) \\ \vdots & & \vdots \\ x(N_2 - 1) & \cdots & x(N_2 - K) \end{bmatrix},$$

where the number of samples N is given by $N = N_2 - N_1 + 1$ with N_1 being the starting sample and N_2 the ending. In this example, we consider not only the unknown parameters to be the AR coefficients $\{a_k\}$ but also the noise variance σ^2. Consequently, our parameter vector is now given by

$$\boldsymbol{\theta} = \begin{bmatrix} a_1 & \dots & a_K & \sigma^2 \end{bmatrix}^T \tag{32}$$

$$= \begin{bmatrix} \mathbf{a}^T & \sigma^2 \end{bmatrix}^T. \tag{33}$$

For $N \gg K$, the weighting matrix associated with this parametrization is given by[37]

$$\mathbf{W} = \frac{N}{\tilde{\sigma}^2} \begin{bmatrix} \widetilde{\mathbf{R}} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\tilde{\sigma}^2} \end{bmatrix}, \tag{34}$$

where $\mathbf{R}$ is covariance matrix for $\mathbf{x}$ and $\widetilde{\mathbf{R}}$ its estimate (we will return to this shortly). It can be observed from (34) that the AR coefficients and the noise variance estimate are independent and can be treated separately. This may sound somewhat curious as the noise variance estimate obviously depends on the estimated AR coefficients. The weighting matrix in (34) leads to the following metric:

$$J = \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \mathbf{W} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \tag{35}$$

$$= \frac{N}{\tilde{\sigma}^2} \left(\begin{bmatrix} \tilde{\mathbf{a}} \\ \tilde{\sigma}^2 \end{bmatrix} - \begin{bmatrix} \mathbf{a} \\ \sigma^2 \end{bmatrix} \right)^T \tag{36}$$

$$\times \begin{bmatrix} \widetilde{\mathbf{R}} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\tilde{\sigma}^2} \end{bmatrix} \left(\begin{bmatrix} \tilde{\mathbf{a}} \\ \tilde{\sigma}^2 \end{bmatrix} - \begin{bmatrix} \mathbf{a} \\ \sigma^2 \end{bmatrix} \right).$$

Writing this out, we obtain

$$J = \frac{N}{\tilde{\sigma}^2} \left(\tilde{\mathbf{a}} - \mathbf{a}\right)^T \widetilde{\mathbf{R}} \left(\tilde{\mathbf{a}} - \mathbf{a}\right) + \frac{N \left(\tilde{\sigma}^2 - \sigma^2\right)^2}{2\tilde{\sigma}^4}. \tag{37}$$

As before, the question remains which estimator to use. Under certain conditions, the well-known least-squares method with appropriately chosen N_1 and N_2 will result in estimates that are asymptotically distributed according to (5) for both the AR coefficients and the noise variance estimates[37,38]. As for the covariance matrix $\mathbf{R}$, the usual sample covariance matrix is well-known to be the maximum likelihood estimator for Gaussian signals[38] and can hence be used in lieu of $\widetilde{\mathbf{R}}$. Note that even if the signal is not Gaussian, the Gaussian assumption is still optimal in a min-max sense when nothing is known about the true pdf[39].

We note in passing that for related parametriziations[17,40], like the reflection coefficients, cepstral coefficients, log area ratios (LAR), or line spectral frequencies (LSF), the Fisher information matrix can be obtained by transforming the Fisher information matrix of the auto-regressive coefficients, akin to the transformation of the sensitivity matrix[18,41] for cepstral coefficients and LSF. For the case of reflection coefficient, a recursive formula for determining the Fisher information matrix exists[42].

IV. EXPERIMENTAL RESULTS

A. Methodology

In the sections to follow, we will present some simulations results. The aim of the first simulations is to demonstrate how the loss incurred by the use of an intermediate pametric description in the VQ process is minimized using the proposed methodology. For each experiment we will among others compare to a naive approache ignoring the different uncertainties associated with the intermediate parameters, corresponding to using $\mathbf{W} = \mathbf{I}$. This results in a least-squares (LS) estimator in the vector quantization process. Additionally, we will compare to the upper bound performance obtained as follows: for each codebook entry, a signal is reconstructed and the 2-norm of the error between this signal and the observed signal is measured. The codebook entry that leads to the lowest error is then chosen as the estimate. This approach, which we refer to as analysis-by-synthesis (AbS) is optimal in the sense that it chooses the codebook entry that best explains the observed signal (and

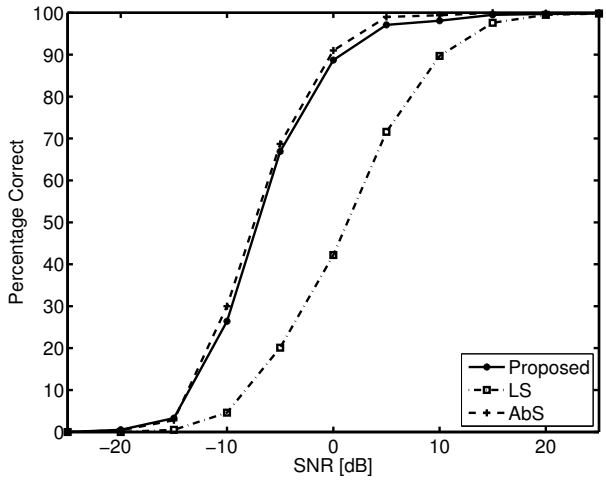

FIG. 1. Results for the sinusoidal model: Percentage of correctly estimated codebook entries as a function of the SNR for $N = 50$.

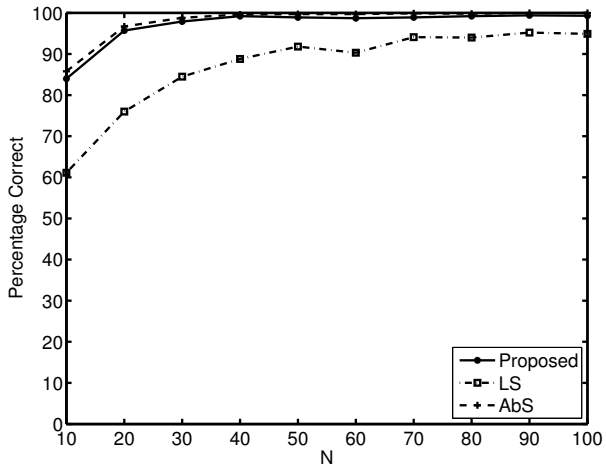

FIG. 2. Results for the sinusoidal model: Percentage of correctly estimated codebook entries as a function of the number of samples N for an SNR of 10 dB.

reconstructs it the best). In fact, it is an implementation of (2) for white Gaussian noise. It is, however, also computationally expensive as it measures distances of N-dimensional signals rather than the M-dimensional parameter vectors. The experiments are carried out by generating a signal $\mathbf{x}$ from a set of parameters from the codebook after which noise is added. In the experiments to follow, the performance is measured as the percentage of correctly estimated codebook entries. For each reported data point, 1000 Monte Carlo trials

are run. If, as has been hypothesized, the proposed metrics are good, it should lead to better estimates than the simple least-squares estimates and to estimates close to those obtained with the AbS method. We note that the estimation errors in the intermediate parameters may cause the selection of a wrong codebook entry when the noise is so large that it causes the estimated parameters to be far from the true parameters, which, in turn, causes the selection of a wrong codebook entry. We note that since a high number of Monte Carlo trials is required to determine the percentages at the desired accuracy, we keep the dimensionality of parameter sets, the codebooks sizes and the number of samples N moderate in size to keep the complexity at a reasonable level. The chosen parameter sets are, however, still within the same order of magnitude as those commonly employed in the literature.

B. Sinusoidal Model

We will now report the results obtained for the various models, and we will start out with the sinusoidal model. For the sinusoidal model, the intermediate parameters are then found using a 8192 point FFT periodogram estimate and these are then quantized using the different metrics. We here use the model, the metric, and the estimators discussed earlier. Moreover, we use a random codebook of size 4096 which has been populated by realization of uniformly distributed phases and frequencies between 0 and 2π and Rayleigh distributed amplitudes. The results are depicted in Figures 1 and 2 as functions of the signal-to-noise ratio (SNR) with $N = 50$ and the number of samples N with $SNR = 0$ dB, respectively. The SNR is here defined as $10 \log_{10} A_1/\sigma^2$. The figures show that the proposed metric outperforms least-squares in the regions of interest. It can be seen from Figure 1 that for extremely low and high SNRs, the performance of the methods tend to 0 and 100 %, respectively, and similar conclusions hold for low and high N, although either extreme is not always achieved for all methods for the region shown. This means that if the estimation error is sufficiently small, which is the case for high SNRs and high N, the correct codebook entry will be chosen, and how small it has to be depends on the method in question. It can

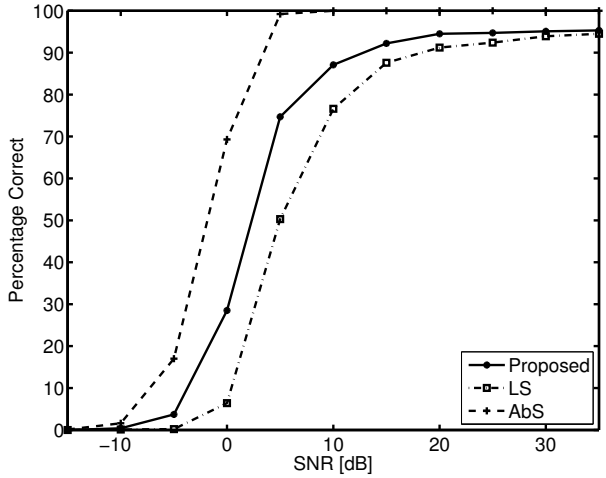

FIG. 3. Results for the harmonic model: Percentage of correctly estimated codebook entries as a function of the SNR for $N = 100$.

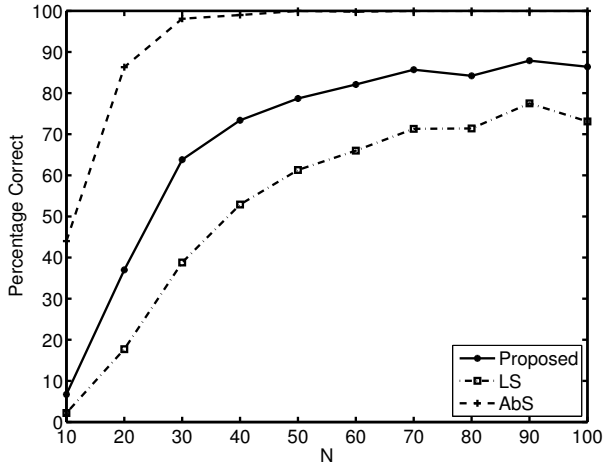

FIG. 4. Results for the harmonic model: Percentage of correctly estimated codebook entries as a function of the number of samples N for an SNR of 10 dB.

also be seen that the proposed metric leads to results that are very close to those obtained using the AbS approach, and this is done at a significantly reduced computation time.

C. Harmonic Model

Proceeding now with reporting the results for the harmonic model, the experimental setup is as follows: the fundamental frequency is found using the so-called approximate

nonlinear least-squares method[34], which is asymptotically efficient, where after the amplitudes and phases are found using approximate least-squares. Note that these estimators are suboptimal for small N and are bound to cause errors. We do, however, use them anyway to provide realistic results. We will use $L = 5$, i.e., five harmonics throughout the experiment, and, as before, we use a randomly generated codebook for generating the observed signal. In this case, the fundamental frequency is generated uniformly on the interval $(0, 2\pi/L]$, phases on the interval $(0, 2\pi]$ and with Rayleigh distributed amplitudes. For this experiment, we use a codebook of size 2048. In each iteration of the Monte Carlo simulations, a new realization is picked and white Gaussian noise is added. In this case, the factor determining the performance, aside from the number of samples N, is the SNR, which for this model is given by

$$SNR = 10 \log_{10} \frac{\sum_{l=1}^{L} A_l^2 l^2}{\sigma^2} \text{ [dB]}. \tag{38}$$

The results are shown in Figures 3 and 4, with the first figure showing the performance as a function of the SNR for $N = 100$ and the second figure showing the performance as a function of the number of samples, N, for an SNR of 10 dB. The results and conclusions are similar to those of the previous experiment, only there is now a sub-optimality gap between the performance of the proposed method and the AbS method. This can be attributed to the use of suboptimal estimators in finding the intermediate parameters. The proposed method is, though, still better than the straightforward least-squares approach and this demonstrates the validity of the approach.

D. Auto-Regressive Model

The final model, for which we will report results, is the auto-regressive one. In the experiments, we employ a codebook trained from 10 different speakers with two utterances of approximately 5 s length for each speaker, all from the Danish EUROM.1 database. Auto-regressive parameters of a 10th order model were extracted from 20 ms segments (non-speech segments were ignored) and a codebook of size 1024 was trained using the LBG algorithm[43].

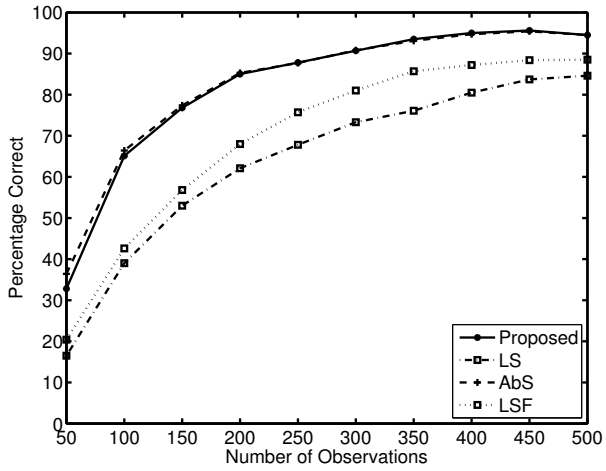

FIG. 5. Results for the auto-regressive model: Percentage of correctly estimated codebook entries as a function of the number of samples N.

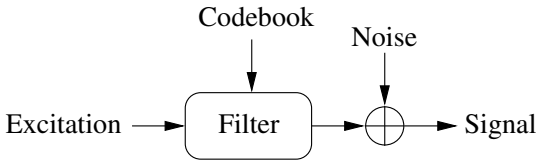

FIG. 6. Model used for generating the test signals for the auto-regressive model.

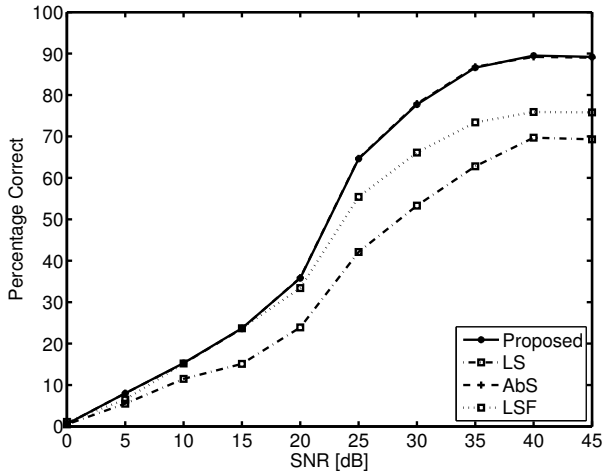

FIG. 7. Results for the auto-regressive model: Percentage of correctly estimated codebook entries as a function of the SNR for $N = 250$.

As the codebook is kept fixed for all methods, uniform weighting is used in the training so as to not favor any particular method. Note that the quality of the codebook is not of

vital importance here, as ability to find the right entries in it is the subject of the present study. Aside from comparing to the AbS and LS methods, as before, we also compare the proposed method to a method employing LSFs[17]. To keep the experiment fair, all methods operate on the same codebook and that the LSFs are one-to-one mappings of each entry in the codebook. Moreoever, it should be stressed that the so-called analysis-by-synthesis method does not measure differences in the parameter space and hence does not suffer from this problem; indeed it would perform exactly the same if LSF or some other factorization were used since these are one-to-one mappings. The experiments are then carried out as follows: an entry in the codebook is picked randomly, a signal is synthesized by filtering white Gaussian noise using the auto-regressive parameters and, from this signal, the parameters are again estimated using the autocorrelation method[38]. Then, the various factorizations and metrics are used for finding the codebook entry that is the best match to the estimated parameters. Here it should be stressed that the accuracy at which it is possible to determine the parameters depends on the nature of $\mathbf{R}$. The results are shown as a function of the number of samples N in Figure 5. It can be seen that, as expected, direct quantization of the parameters without appropriate weighting performs the poorest, as expected. Then follows the method employing LSFs. It can be seen that the proposed method performs almost as well as the AbS method, which means that the method is close to the optimal performance for this task. It can also generally be observed that all the methods improve as a function of N as the quality of the involved estimates improves. In the next experiment, everything will be carried out as before, except that the assumptions under which the proposed method was derived will be violated by an additional, additive white Gaussian noise source, as depicted in Figure 6. The performance of the methods will then be observed as a function of the SNR between the variance of the output of the filter and that of the noise source. The results are shown in Figure 7. The general conclusions are the same as for the previous experiment: the LS approach performs the worst, then the LSF-based method follows, and the proposed approach performs similarly to the AbS approach. Interestingly, it can be observed that as the SNR is worsened, the performance of the various methods appear to perform similarly,

with the LSF, AbS and the proposed method having almost equal performance for SNRs below 20 dB. Note that it may be difficult to distinguish the last two methods as the curves coincide. The general conclusion is that the proposed method performed essentially without loss compared to the optimal estimator, the AbS method, and this at a significantly reduced computational complexity.

E. Weighting Matrix

In deriving the closed-form expression for the weighting matrix, it was assumed that the parameter estimates will be distributed according to the Fisher information matrix. It remains then to quantify whether this is actually the case, and we will address this in our final experiment. In this experiment, a random realization of a parameter set is used to compute the weighting matrix for the various models in Section III. Then, the observed signal is generated from those parameter sets and noise is added according to the respective models. The sample covariance matrix of the estimation errors obtained with the various estimators used in these experiments is then computed over 1000 realizations for each data point. In theory, the so-obtained covariance matrices should tend to the optimal weighting matrix as N is increased, however, since suboptimal estimators are used, their finite sample length performance cannot be known beforehand. Consequently, we measure the distance between the two matrices using the Frobenius normalized by the size of the matrix squared resulting in a mean squared error. Note that since the matrices may differ by a constant factor without affecting the result, the optimal scaling that minimizes the Frobenius norm of the difference between the two is found and applied for each pair of matrices. The results are shown in Figure 8 as N is increased. It can be seen that for all the three models used and their estimators, the estimation errors tend toward the closed-form expression for the weighting matrix, albeit computed for the true parameters instead of the estimated ones. In this connection we note that the relation between the weighting matrix obtained for the true and estimated parameters, respectively, does not lend itself to a simple investigation

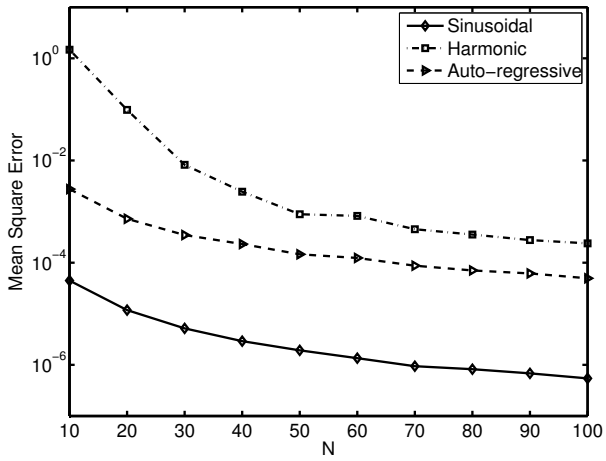

FIG. 8. Mean squared error between the weighting matrix for the true parameters and the sample covariance matrix for the estimation errors for the three models.

due to the complicated manner in which they converge[29].

V. CONCLUSIONS AND FUTURE WORK

In this paper, metrics for VQ-based speech enhancement and separation have been proposed. The metrics were derived based on statistical arguments and expressions for the asymptotic distribution of maximum likelihood estimators. It essentially takes the uncertainties and dependencies of different parameters into account in the quantization process. This was then demonstrated to lead to superior estimates in Monte Carlo simulations with a vector quantizer as compared to the commonly used squared error measure. Moreover, the proposed metrics perform close to the optimal performance (in the sense of maximum likelihood) under good conditions like a high number of samples and high signal-to-noise ratios, showing that only a small loss is incurred by operating on intermediate parameters in the process. It remains to be seen, of course, firstly, whether it is possible to obtain such simple closed-form expressions for the many different signal models employed in speech processing, aside from the few treated here, and, secondly, how these differ from the ones currently in use. Moreover, an interesting question is whether and, if so, how the proposed

methodology can be extended to account for the quality of the reconstructed signal using state-of-the-art perceptual measures, rather than finding the most likely explanation for the observed signal.

References

[1] E. Zavarehei, S. Vaseghi, and Q. Yan, "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing", IEEE Trans. Audio, Speech, and Language Process. **15**, 1194–1203 (2007).

[2] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement", IEEE Trans. Audio, Speech, and Language Process. **14**, 163–176 (2006).

[3] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments", IEEE Trans. Audio, Speech, and Language Process. **15**, 441–452 (2007).

[4] M. Kuropatwinski and W. B. Kleijn, "Estimation of the excitation variances of speech and noise AR-models for enhanced speech coding", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 1, 669–672 (2001).

[5] T. V. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement", IEEE Trans. Audio, Speech, and Language Process. **4**, 383–389 (1996).

[6] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising", in *EUROSPEECH*, 1009–1012 (2003).

[7] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition", Elsevier Computer Speech and Language **24**, 30–44 (2010).

[8] D. P. W. Ellis and R. J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing **5**, 957–960 (2006).

[9] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation", EURASIP J. on Advances in Signal Processing **1**, 1–15 (2007).

[10] P. Mowlaee, M. G. Christensen, and S. H. Jensen, "New results on single-channel speech separation using sinusoidal modeling", IEEE Trans. on Audio, Speech and Language Processing **19(5)**, 1265–1277 (2011).

[11] Y. Ephraim, "Statistical-model-based speech enhancement systems", Proc. IEEE **80**, 1526–1555 (1992).

[12] B. S. Atal and M. R. Schroeder, "Optimizing predictive coders for minimum audible noise", IEEE Trans. Acoust., Speech, Signal Process. **27(3)**, 247–254 (1979).

[13] M. R. Schroeder, B. S. Atal, and J. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear", J. Acoust. Soc. Am. **66(6)**, 1647–1652 (1979).

[14] A. H. Gray Jr. and J. D. Markel, "Distance measures for speech processing", IEEE Trans. Acoust., Speech, Signal Processing **24(5)**, 380–391 (1976).

[15] R. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing", IEEE Trans. Acoust., Speech, Signal Process. **28**, 367–376 (1980).

[16] B.-H. Juang, D. Wong, and A. H. Gray, Jr., "Distortion performance of vector quantization for LPC voice coding", IEEE Trans. Acoust., Speech, Signal Process. **30**, 294–304 (1982).

[17] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame", IEEE Trans. Speech and Audio Processing **1**, 3–14 (1993).

[18] W. R. Gardner and B. D. Rao, "Theoretical analysis of the high-rate vector quantization of lpc parameters", IEEE Trans. Speech Audio Process. **3(5)**, 367–381 (1995).

[19] J. D. Johnston, "Transform coding of audio signal susing perceptual noise criteria", IEEE J. Select. Areas Commun. 314–323 (1988).

[20] R. Vafin and W. B. Kleijn, "Entropy-constrained polar quantization: Theory and an application to audio coding", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 2, 1837–1840 (2002).

[21] P. Korten, J. Jensen, and R. Heusdens, "High-resolution spherical quantization of sinusoidal parameters", IEEE Transactions on Audio, Speech, and Language Processing **15**, 966–981 (2007).

[22] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration", EURASIP J. on Advances in Signal Processing **9**, 1292–1304 (2004).

[23] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. i. model structure", J. Acoust. Soc. Am. **99(6)**, 3615–3622 (1996).

[24] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. ii. simulations and measurements", J. Acoust. Soc. Am. **99(6)**, 3623–3631 (1996).

[25] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis* (Addison-Wesley, Reading, Massachusetts) (1991), pp. 1-524.

[26] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall Signal Processing Series (Prentice Hall PTR, Upper Saddle River, New Jersey) (1993), pp. 1-595.

[27] P. Stoica and T. Söderström, "On reparametrization of loss functions used in estimation and the invariance principle", Elsevier Signal Processing **17**, 383–387 (1989).

[28] A. L. Swindlehurst and P. Stoica, "Maximum likelihood methods in radar array signal processing", Proc. IEEE **86(2)**, 421–441 (1998).

[29] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules", IEEE SP Mag. **21(4)**, 36–47 (2004).

[30] P. M. Djuric, "Asymptotic MAP criteria for model selection", IEEE Trans. Signal Processing **46(10)**, 2726–2735 (1998).

[31] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: Survey, new results and an application", IEEE Trans. Signal Processing **48(2)**, 338–352 (2000).

[32] M. G. Christensen, A. Jakobsson and S. H. Jensen, "Joint high-resolution fundamental

frequency and order estimation", IEEE Trans. on Audio, Speech and Language Processing **15(5)**, 1635–1644 (2007).

[33] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement", IEEE Trans. Acoust., Speech, Signal Processing **34(5)**, 1124–1138 (1986).

[34] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, volume 5 of *Synthesis Lectures on Speech & Audio Processing* (Morgan & Claypool Publishers, San Rafael, California) (2009), pp. 1-160.

[35] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals", Signal Processing **80**, 1937–1944 (2000).

[36] P. P. Vaidyanathan, *The Theory of Linear Prediction*, volume 2 of *Synthesis Lectures on Signal Processing* (Morgan & Claypool Publishers, San Rafael, California) (2007), pp. 1-184.

[37] S. M. Kay, *Modern Spectral Estimation: Theory and Application*, Prentice Hall Signal Processing Series (Prentice Hall, Upper Saddle River, New Jersey) (1988), pp. 1-539.

[38] P. Stoica and R. Moses, *Spectral Analysis of Signals* (Pearson Prentice Hall, Upper Saddle River, New Jersey) (2005), pp. 1-452.

[39] P. Stoica and P. Babu, "The Gaussian data assumption leads to the largest Cramér-Rao bound [lecture notes]", IEEE Signal Process. Mag. **28**, 132–133 (2011).

[40] W. B. Kleijn and K. K. Paliwal, "Quantization of LPC Parameters", in *Speech Coding and Synthesis*, edited by W. B. Kleijn and K. K. Paliwal, chapter 12 (Elsevier Science B.V.) (1995).

[41] J. Samuelsson and P. Hedelin, "Recursive coding of spectrum parameters", IEEE Trans. Speech Audio Process. **9**, 492–503 (2001).

[42] S. Kay and J. Makhoul, "On the statistics of the estimated reflection coefficients of an autoregressive process", IEEE Trans. Acoust., Speech, Signal Process. **31**, 1447–1455 (1983).

[43] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Communications and Information Theory (Kluwer Academic Publishers, Boston/Dordrecht/London)

(1993), pp. 1-732.

List of Figures

FIG. 1 Results for the sinusoidal model: Percentage of correctly estimated codebook entries as a function of the SNR for $N = 50$. 16

FIG. 2 Results for the sinusoidal model: Percentage of correctly estimated codebook entries as a function of the number of samples N for an SNR of 10 dB. . . . 16

FIG. 3 Results for the harmonic model: Percentage of correctly estimated codebook entries as a function of the SNR for $N = 100$. 18

FIG. 4 Results for the harmonic model: Percentage of correctly estimated codebook entries as a function of the number of samples N for an SNR of 10 dB. . . . 18

FIG. 5 Results for the auto-regressive model: Percentage of correctly estimated codebook entries as a function of the number of samples N. 20

FIG. 6 Model used for generating the test signals for the auto-regressive model. . . 20

FIG. 7 Results for the auto-regressive model: Percentage of correctly estimated codebook entries as a function of the SNR for $N = 250$. 20

FIG. 8 Mean squared error between the weighting matrix for the true parameters and the sample covariance matrix for the estimation errors for the three models. 23